



# UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE  
United States Patent and Trademark Office  
Address: COMMISSIONER FOR PATENTS  
P.O. Box 1450  
Alexandria, Virginia 22313-1450  
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
-----------------	-------------	----------------------	---------------------	------------------

10/626,984

07/25/2003

Kenji Kawai

013.0207.US.UTL

4146

22895

7590

11/26/2008

CASCADIA INTELLECTUAL PROPERTY  
500 UNION STREET  
SUITE 1005  
SEATTLE, WA 98101

EXAMINER

HARPER, LEON JONATHAN

ART UNIT

PAPER NUMBER

2166

MAIL DATE

DELIVERY MODE

11/26/2008

PAPER

**Please find below and/or attached an Office communication concerning this application or proceeding.**

The time period for reply, if any, is set in the attached communication.



## **DETAILED ACTION**

### ***Response to Amendment***

1. The amendment filed 9/2/2008 has been entered. Claims 1,9,17,18,20,35,37,52 and 53 have been amended. Claims 4 and 12 have been presently cancelled. Accordingly, claims 1, 3,5,6,9,11,13,14, 17-27, 35-44, 48,52 and 53 are pending in this office action.

### ***Allowable Subject Matter***

Claims 19-23, 36-40 are objected to as being dependent upon a rejected base claim, but would be allowable if rewritten in independent form including all of the limitations of the base claim and any intervening claims.

### ***Claim Rejections - 35 USC § 103***

The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

The factual inquiries set forth in *Graham v. John Deere Co.*, 383 U.S. 1, 148

USPQ 459 (1966), that are applied for establishing a background for determining obviousness under 35 U.S.C. 103(a) are summarized as follows:

1. Determining the scope and contents of the prior art.
2. Ascertaining the differences between the prior art and the claims at issue.
3. Resolving the level of ordinary skill in the pertinent art.

Art Unit: 2166

4. Considering objective evidence present in the application indicating obviousness or nonobviousness.

Claims 1-6, 9-14, 17-23 ,35-40, 52,53 are rejected under 35 U.S.C. 103(a) as being unpatentable over Wo 03060766 (hereinafter Lind) (art of record) in view of US 6560597 (hereinafter Dhil).

As for claim 1 Lind discloses: a scoring module determining a score which is assigned to at least one concept that has been extracted from a plurality of electronically-stored documents (See page 17 lines 20-24 note: definition of document corpus) wherein the score is calculated as a function of a summation of a frequency of occurrence of the at least one concept within at least one such document, a concept weight based on a number of terms for the at least one concept, a structural weight, and a corpus weight, forming the score assigned to the at one concept as a normalized score vector for each such document, and determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector (See page 30 line 30- page 31 line 1).

; (See page 7 lines 20-24) a clustering module forming clusters of the documents comprising a selection sub module selecting a set of candidate seed documents from the plurality of documents; a seed document identification submodule identifying a set of seed documents by applying the similarity to each such candidate seed document and selecting those candidate seed documents that are sufficiently unique from other candidate seed documents as the seed documents; a non-seed document identification submodule identifying a plurality of non-seed documents; a

Art Unit: 2166

comparison submodule determining the similarity between each non-seed document and a center of each cluster; and a clustering submodule grouping each such non-seed document into a cluster with a best fit, subject to a minimum fit See page 28 line 8-16 note representative= seed).by evaluating the score for the at least one concept of each document for a best fit to the clusters and assigning each document to the cluster with the best fit; and (See page 19 lines 4-10). While Lind does not differ substantially from the claimed invention the disclosure of a threshold module determining the similarity between each of the documents grouped into each cluster based on the center of the cluster and the scores assigned to each of the at least one concepts in that document dynamically determining a threshold for each cluster as a function of the similarity between each of the documents, and identifying and reassigning each of the documents having the similarity falling outside the threshold are not necessarily explicit. Dhill however does disclose a threshold module determining the similarity between each of the documents grouped into each cluster based on the center of the cluster and the scores assigned to each of the at least one concepts in that document dynamically determining a threshold for each cluster as a function of the similarity between each of the documents, (See column 3 lines 55-60 and column 5 line 55- column 6 line 5) and identifying and reassigning each of the documents having the similarity falling outside the threshold (See column 3 lines 60-65). It would have been obvious to an artisan of ordinary skill in the pertinent art at the time the invention was made to have incorporated the teaching of Dhill into the system of Lind. The modification would have been obvious because the two references are concerned with the solution to problem of efficient

Art Unit: 2166

document scoring and clustering, therefore there is an implicit motivation to combine these references. In other words, the ordinary skilled artisan, during his/her quest for a solution to the cited problem, would look to the cited references at the time the invention was made. Consequently, the ordinary skilled artisan, would have been motivated to combine the cited references since Dhill's teaching would enable Lind's users to reclassify documents based on the center of the cluster..

As for claim 3 the rejection of claim 1 is incorporated, and further Lind discloses: a compression module compressing the score through logarithmic compression (See page 17 line 30-34).

As for claim 4 the rejection of claim 1 is incorporated, and further Lind discloses: the scoring module calculating the concept weight as a function of a number of terms comprising the at least one concept (See page 21 lines 25-28).

As for claim 5 the rejection of claim 1 is incorporated, and further Lind discloses: the scoring module calculating the structural weight as a function of a location of the at least one concept within the at least one such document (See page 18 lines 10-14).

As for claim 6 the rejection of claim 1 is incorporated, and further Lind discloses: the scoring module calculating the corpus weight as a function of a reference count of

Art Unit: 2166

the at least one concept over the plurality of documents (See page 18 lines 19- 21 note: this is an inverse weight of the reference count).

Claims 9, 11-14 are method claims corresponding to system claims 1, 3-6 respectively, and are thus rejected for the reasons set forth in the rejection of claims 1, 3-6.

Claim 17 is rejected for the same reasons as claim 9.

As for claim 18 Lind discloses: a scoring module scoring a document in an electronically-stored document set comprising: a frequency module determining a frequency of occurrence of at least one concept within a document (See page 18 lines 1-3); and a concept weight module analyzing a concept weight reflecting a specificity of meaning for the at least one concept within the document wherein the concept weight is based on a number of terms for the at least one concept (See page 25 lines 27-30 note:  $rtc(t,c)$  is a value based on meaning); a structural weight module analyzing a structural weight reflecting a degree of significance based on structural location within the document for the at least one concept (See page 18 lines 8-13), a corpus weight module analyzing a corpus weight inversely weighing a reference count of occurrences for the at least one concept within the document (See page 18 lines 19- 21 note: this is an inverse weight of the reference count); and a scoring evaluation module evaluating a

Art Unit: 2166

score to be associated with the at least one concept as a function of the frequency, concept weight, structural weight, and corpus weight; (See page 21 24-27) and

A vector module forming the score assigned to the at least one concept as a normalized score vector for each such document in the electronically-stored document set, and a determination module determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector (See page 30 line 30- page 31 line 1). A clustering module grouping the documents by the score into a plurality of clusters comprising; a selection submodule evaluating a set of candidate seed documents selected from the electronically-stored document set; a cluster seed submodule identifying seed documents by applying the similarity to each such candidate seed document and selecting those candidate seed documents that are sufficiently unique from other candidate seed documents as the seed documents; an identification submodule identifying a plurality of non-seed documents; a comparison submodule determining the similarity between each non-seed document and a center of each cluster; and a clustering submodule assigning each non-seed document to the cluster with the best fit, subject to a minimum fit (See page 28 line 8-16 note representative= seed, (See column 5 lines 35-42).

While Lind does not differ substantially from the claimed invention the disclosure of a threshold module relocating outlier documents, comprising determining the similarity between each of the documents groups into each cluster based on the center of the cluster and the scores assigned to each of the at least one concepts in that document , dynamically determining a threshold for each cluster as a function of the



Art Unit: 2166

similarity between each of the documents, and identifying and reassigning each of the documents with the similarity falling outside the threshold are not necessarily explicit. Dhill however does disclose a threshold module determining the similarity between each of the documents grouped into each cluster based on the center of the cluster and the scores assigned to each of the at least one concepts in that document dynamically determining a threshold for each cluster as a function of the similarity between each of the documents (See column 3 lines 55-60 and column 5 line 55- column 6 line 5); and identifying and reassigning each of the documents having the similarity falling outside the threshold (See column 3 lines 60-65). It would have been obvious to an artisan of ordinary skill in the pertinent art at the time the invention was made to have incorporated the teaching of Dhill into the system of Lind. The modification would have been obvious because the two references are concerned with the solution to the problem of efficient document scoring and clustering, therefore there is an implicit motivation to combine these references. In other words, the ordinary skilled artisan, during his/her quest for a solution to the cited problem, would look to the cited references at the time the invention was made. Consequently, the ordinary skilled artisan, would have been motivated to combine the cited references since Dhill's teaching would enable Lind's users to reclassify documents based on the center of the cluster..

Claims 35, is a method claim comprising substantially the same limitation as system claims 18, and are thus rejected for the reasons set forth in the rejection of claim 18.

Claim 52 is rejected for substantially the same reasons as claim 35,

Claim 53 is an apparatus claim corresponding to method claim 18 and is thus rejected for the same reasons as claim 18.

Claims 24-27 and 41-44 and claims is rejected under 35 U.S.C. 103(a) as being unpatentable over Lind and Dhill as applied to claim 18 and 35 above, and further in view of US 6675159 (hereinafter Klein) (art of record)

As for claim 24 the rejection of claim 18 is incorporated, and further Klein discloses: a global stop concept vector cache maintaining concepts and terms (See column 18 lines 17-20 and See column 14 lines 45-49); and a filtering module filtering selection of the at least one concept based on the concepts and terms maintained in the global stop concept vector cache (See column 14 lines 45-50). It would have been obvious to an artisan of ordinary skill in the pertinent art at the time of the invention to have incorporated the teachings of Klein into the system of Lind. The modification would have been obvious because queries and documents are linked in the fact that words are

Art Unit: 2166

the entities that are being processed. Therefore, any transformation capable of being made to a query should be able to applied to documents too, this makes all document management systems more efficient and easier to maintain.

As for claim 25 the rejection of claim 18 is incorporated, and further Klein discloses: a parsing module identifying terms within at least one document in the document set, and combining the identified terms into one or more of the concepts (See column 2 lines 53-56).

As for claim 26 the rejection of claim 25 is incorporated, and further Klein discloses: the parsing module structuring each such identified term in the one or more concepts into canonical concepts comprising at least one of word root, character case, and word ordering (See column 14 lines 63-67).

As for claim 27 the rejection of claim 25 is incorporated, and further Klein discloses wherein at least one of nouns, proper nouns and adjectives are included as

Claims 41-44, are method claims corresponding to system claims 24-27, respectively and are thus rejected for the same reasons as set forth in the rejection of claims 24-27,.

Art Unit: 2166

Claims 31 and 48 are rejected under 35 U.S.C. 103(a) as being unpatentable over Lind and Dhill as applied to claim 29 above, and further in view of Klein.

As for claim 31, the rejection of claim 30 is incorporated, and further Klein discloses: the similarity submodule calculating the similarity in accordance with the formula

$$\cos \theta_{ab} = \frac{(S_s \cdot S_b)}{S_a S_b}$$

Where  $\cos \theta_{ab}$  comprises a similarity between a document A and a document B,  $S_a$  comprises a score vector for document A and  $S_b$  comprises a score vector for document B.

Claim 48 is a method claim corresponding to the system of claim 31 respectively and is thus rejected for the same reasons as set forth in the rejection of claim 31.

### ***Response to Arguments***

Applicant's arguments filed 9/2/08 have been fully considered but they are not persuasive.

#### **Applicant argues:**

Lindh teaches determining a term weight for each unique word in a text (Lindh, p. 17, lines 20-23). The term weight is calculated using a TFIDF equation, which includes multiple parameters, including a number of occurrences of a term in a document, a total number of terms in the document, a number of documents in which the term exists, a total number of documents in the document corpus, and a weight function dependent on the positions of the terms in the document (Lindh, p. 17, line 30-p. 18, line 9). The parameters are then entered into the TFIDF equation to calculate the weight for a particular term. Thus, a number of occurrences of a particular terms are determined for a single document, as well as a total number of terms included in the document. Therefore, Lindh fails to determine a number of terms for a concept as a concept weight. Lindh also fails to teach a score, which is assigned to at least one concept. Instead, Lindh teaches calculating a relation value for a given term and a given concept (Lindh, p. 22, line 34-p. 23, line 3). The relation value is determined using a given equation, which is based on a term weight and a document-concept relationship value (Id.). The term weight is calculated using the TFIDF equation (Lindh, p. 17, line 30-p. 18, line 9) and the document-concept relationship value describes a relationship between a document and a concept (p. 23, lines 7-9). Thus, the relation value fails to consider a concept weight, which is based on a number of terms for a concept.

Art Unit: 2166

Therefore, Lindh teaches a relation value for a given term and given concept, rather than a score that is calculated as a function of a summation of a frequency of occurrence of at least one concept within at least one such document, a concept weight based on a number of terms for the at least one concept, a structural weight, and a corpus weight, per Claims 1, 9, 17, 18, 35, 52, and 53.

**Examiner responds:**

Examiner is not persuaded. Examiner is entitled to give claim limitations their broadest reasonable interpretation in light of the specification. Interpretation of Claims- Broadest Reasonable Interpretation: During patent examination, the pending claims must be 'given the broadest reasonable interpretation consistent with the specification.' Applicant always has the opportunity to amend the claims during prosecution and broad interpretation by the examiner reduces the possibility that the claim, once issued, will be interpreted more broadly than is justified. In re Prater, 162 USPQ 541,550-51 (CCPA 1969). In this case applicant argues that Lindh fails to determine a number of terms for a concept as a concept weight however a concepts are as broad thoughts or ideas. Moreover concept's are abstract so when the equation calculates a weight for a term that is also a concept weight. Accordingly, Lindh does in fact disclose concepts and concept weights.

Art Unit: 2166

**Applicant argues:**

Lindh teaches applying a clustering algorithm to a document corpus, rather than selecting a set of candidate seed documents from a plurality of documents. Further, Lindh fails to teach identifying a set of seed documents from the set of candidate seed documents. As described above, Lindh applies a clustering algorithm, such as a k-means algorithm to a document corpus to remove documents that are similar. After the clusters have been identified, a representative document vector is determined and assigned to each cluster (Lindh, p. 28, lines 20-23). Next, the representative document vector is added to the cluster, and documents belonging to the cluster are removed except for the representative document vector (Lindh, p. 28, lines 16-24; FIGURE 9A). As the clustering algorithm is applied to the complete document corpus, a set of candidate seed documents are not selected, nor is a set of seed documents identified based on a similarity determined for each document. Thus, Lindh teaches applying a clustering algorithm to a document corpus to identify clusters of the documents, rather than identifying a set of seed documents by applying the similarity to each such candidate seed document in each category and selecting those candidate seed documents that are sufficiently unique as the seed documents.

**Examiner responds:**

Examiner is not persuaded. Examiner is entitled to give claim limitations their broadest reasonable interpretation in light of the specification. Interpretation of Claims- Broadest Reasonable Interpretation: During patent examination, the pending claims

Art Unit: 2166

must be 'given the broadest reasonable interpretation consistent with the specification.'

Applicant always has the opportunity to amend the claims during prosecution and broad interpretation by the examiner reduces the possibility that the claim, once issued, will be interpreted more broadly than is justified. In re Prater, 162 USPQ 541,550-51 (CCPA 1969). A representative document is another type of seed document since representative document vector is based on similarity in the documents.



**Conclusion**

**THIS ACTION IS MADE FINAL.** Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

A shortened statutory period for reply to this final action is set to expire **THREE MONTHS** from the mailing date of this action. In the event a first reply is filed within **TWO MONTHS** of the mailing date of this final action and the advisory action is not mailed until after the end of the **THREE-MONTH** shortened statutory period, then the shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than **SIX MONTHS** from the mailing date of this final action.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Leon J. Harper whose telephone number is 571-272-0759. The examiner can normally be reached on 7:30AM - 4:00Pm.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Hosain T. Alam can be reached on 571-272-3978. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Art Unit: 2166

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see <http://pair-direct.uspto.gov>. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

*LJH*

*Leon J. Harper*

*November 23, 2008*

/Hosain T Alam/

Supervisory Patent Examiner, Art Unit 2166